ISSN 1695-5498

HTXT

HIPERTEXT.NET · REVISTA ACADÉMICA SOBRE DOCUMENTACIÓN DIGITAL Y COMUNICACIÓN INTERACTIVA

upf. Universitat Pompeu Fabra Barcelona

Número 28 · Mayo de 2024

Taxonomies and Ontologies in Wikipedia and Wikidata: An In-Depth Examination of Knowledge Organization Systems

Miquel Centelles
Universitat de Barcelona

Núria Ferran-Ferrer
Universitat de Barcelona

Taxonomías y ontologías en Wikipedia y Wikidata: un examen detallado de los sistemas de organización del conocimiento

ABSTRACT

This article examines Wikipedia's knowledge organization system (KOS) and the broader KOS of Wikidata. We study the structure, functions, and relationship of Wikipedia's KOS to concepts like taxonomies and folksonomies, highlighting its unique characteristics compared to social media. A significant aspect of our examination is the gender-related content classification in the Catalan edition of Wikipedia (Viquipèdia), which notably excludes female categories and non-binary gender classifications. We explore the potential implications of these restrictions on gender bias within the platform. Furthermore, we broaden our investigative methodology to assess the KOS of Wikidata. Wikidata is a dataset built on ontological principles, designed to enhance and enrich Wikipedia's digital, collaborative encyclopedia. The findings shed light on the presence or absence of gender bias and contribute to the ongoing discourse on promoting inclusivity and diversity in online knowledge sharing.

RESUMEN

Este artículo examina el sistema de organización del conocimiento (KOS, por sus siglas en inglés) de Wikipedia y el KOS más amplio de Wikidata. Estudiamos la estructura, funciones y la relación del KOS de Wikipedia con conceptos como taxonomías y folksonomías, resaltando sus características únicas en comparación con las redes sociales. Un aspecto significativo de nuestro análisis es la clasificación de contenidos relacionados con el género en la edición catalana de Wikipedia (Viquipèdia), que notablemente excluye categorías de género femenino y clasificaciones de género no binario. Exploramos las posibles implicaciones de estas restricciones en el sesgo de género dentro de la plataforma. Además, ampliamos nuestra metodología de investigación para evaluar el KOS de Wikidata. Wikidata es un conjunto de datos construido sobre principios ontológicos, diseñado para mejorar y enriquecer la enciclopedia digital y colaborativa de Wikipedia. Los hallazgos arrojan luz sobre la presencia o ausencia de sesgo de género y contribuyen al continuo debate sobre la promoción de la inclusividad y la diversidad en el intercambio de conocimientos en línea.

KEYWORDS

Wikipedia; Wikidata; Taxonomy; Ontology; Knowledge Organization System; KOS.

PALABRAS CLAVE

Wikipedia; Wikidata; Taxonomía; Ontología; Sistema de Organización del Conocimiento; KOS.

Centelles, M. & Ferran-Ferrer, N. (2024). Taxonomies and Ontologies in Wikipedia and Wikidata: An In-Depth Examination of Knowledge Organization Systems. *Hipertext.net*, (28), 33-48. https://doi.org/10.31009/hipertext.net.2024.i28.04

Taxonomies i ontologies a Viquipèdia i Wikidata: Un exàmen detallat dels sistemes d’organització del coneixement

RESUM

Aquest article examina el sistema d’organització del coneixement (KOS, per les sigles en anglès) de Wikipedia i el KOS més ampli de Wikidata. S’estudia l’estructura, funcions i la relació del KOS de Wikipedia amb conceptes com taxonomies i folksonomies, ressaltant-ne les característiques úniques en comparació amb les xarxes socials. Un aspecte significatiu de la nostra anàlisi és la classificació de continguts relacionats amb el gènere a l’edició catalana de Wikipedia (Viquipèdia), que notablement exclou categories de gènere femení i classificacions de gènere no binari. Explorem les possibles implicacions d’aquestes restriccions al biaix de gènere dins de la plataforma. A més, ampliem la nostra metodologia de recerca per avaluar el KOS de Wikidata. Wikidata és un conjunt de dades construït sobre principis ontològics, dissenyat per millorar i enriquir l’enciclopèdia digital i col·laborativa de Wikipedia. Les troballes donen llum sobre la presència o absència de biaix de gènere i contribueixen al debat continu sobre la promoció de la inclusivitat i la diversitat en l’intercanvi de coneixements en línia.

PARAULES CLAU

Viquipèdia; Wikipedia; Wikidata; Taxonomia; Ontologia; Sistema d’organització del coneixement; KOS.

1. Introduction

The Wikipedia is the most widely used resource in the educational field, and it ranks among the most frequently visited websites, alongside Google, with more than five billion readers in nearly 300 available languages worldwide (Singer et al., 2017). This digital encyclopedia has transformed the way information is produced and distributed through open collaboration, as anyone can add and edit its content. In fact, it is maintained by a global community of volunteers and, therefore, has the unique potential to facilitate equitable knowledge production based on common goods and the provision of virtual negotiation spaces. It is undoubtedly one of the largest human cooperation efforts ever, both in terms of the number of people involved (hundreds of thousands) and the magnitude of the work created (tens of millions of articles). In fact, the Wikipedia Foundation has a very ambitious goal to be the sum of all existing human knowledge (Ferran-Ferrer et al., 2021).

This article delves into an investigation of Wikipedia’s knowledge organization system, the structure, functions, and the relationship of Wikipedia’s knowledge organization system with concepts such as taxonomies and folksonomies while emphasizing its unique attributes in comparison to social media environments. It poses a case study (Yin, 2009) on the Catalan edition, Viquipèdia, an edition, which notably excludes female categories and non-binary classifications for organizing content (Maciá, 2022). In other Wikipedias, such as the Spanish, English, or French versions, content can be browsed based on gender, as they allow the creation of categories based on gender differentiation.

Categories are linguistic, social, and cultural constructs that present the contents of the encyclopedia. These categories reflect a society with a strong gender bias, and failing to make the presence of articles about women and other marginalized genders visible not only perpetuates this bias but also does not meet the needs of Wikipedia users who consult or edit its content.

From our study, it appears that using Wikidata is the knowledge organization system of Wikipedia, as Wikidata does not exacerbate the gender bias present in society (Zhang and Terveen, 2021). Additionally, Wikidata and its property related to gender identities already label 81.93% of human entities (by July 2023). Considering that Wikipedia plays a pivotal role as a widely utilized learning resource with profound implications for education (Soler-Adillon et al., 2018; Dawe and Robinson, 2017), it is imperative that any invisibilization of women, who represent 50% of the global population, is deemed unacceptable.

2. Wikipedia’s Analysis of the Knowledge

Organization System

2.1. Categorization of Wikipedia

For the purpose of categorization and characterization of the KOS applied on Wikipedia, we position ourselves within the discipline of knowledge organization. Within it, we prioritize the analytical framework proposed by (Zeng, 2008). This author identifies two complementary perspectives: the degree of complexity of the KOS's structure and the functions performed by the KOS.

First, the degree of complexity of the structure oscillates between two extremes: a flat (or one-dimensional) structure and a multidimensional structure. In the first structure, the flat one, concepts do not have any relationship with each other. In the multidimensional structure, concepts are related to each other, either forming hierarchies based on the lesser or greater semantic scope they represent or through semantic associations based on relationships like agent/instrument, cause/effect, and so on.

And second, the functional diversity is also organized in a crescendo, with four milestones: (i) eliminating ambiguity; (ii) managing synonyms or equivalents; (iii) establishing semantic relationships between terms/concepts, particularly hierarchical and associative relationships; and (iv) presenting relationships and properties of concepts in knowledge models.

In most cases, both perspectives are directly proportional. In other words, the more functions a knowledge organization system assumes, the higher its structural complexity.

Regarding the degree of structural complexity, Wikipedia's KOS is composed of categories, which represent the concepts covered in the articles and other types of content in the encyclopedia. These categories are related to each other through a semantic hierarchy, linking categories with a broader or more general meaning to categories with a narrower or more specific meaning. All categories in Wikipedia's KOS are positioned in the hierarchy by connecting to at least one more general category and two or more specific categories. The semantic hierarchical relationship is the only one that fully structures Wikipedia's KOS.

As for the functions performed by Wikipedia's KOS, the categorization guidelines outlined in Wikipedia:Categorització (Wikimedia, 2023f) are clear. Its primary function is to group knowledge by encyclopedia editors and make it more accessible to readers, facilitating navigation between pages. Additionally, the KOS allows for the automatic generation of lists on related topics within a subject.

The comparison between the structure and functions of Wikipedia's KOS and the typology of systems proposed by Zeng (2008) allows us to identify two types to place it within: taxonomies and categorization schemes. The taxonomies are defined as "loosely formed grouping schemes," while the categories are defined as "divisions of items into ordered groups or categories based on particular characteristics." The key distinction lies in the condition of the hierarchy relationship established between categories: loose or light in the case of categorization schemes versus specific or predefined in the case of taxonomies.

In the context of a digital artifact like Wikipedia, where collaborative construction is a hallmark, the two aforementioned perspectives do not fully define the framework of KOS categorization. A third perspective needs to be incorporated, taking into account the decision-making model for the construction of KOS, in line with its structure and functions. In this third perspective of characterizing Wikipedia's KOS, one important principle to consider is that any Wikipedian can initiate the creation of a new category. Quickly, we can associate this possibility with folksonomies, where the foundation of their definition precisely includes "collaborative categorization". Apart from their collaborative nature, folksonomies possess four additional characteristics (Yedid, 2013). They typically originate in digital and web-based settings. These systems rely on the application of uncontrolled, natural language tags, devoid of a structured, non-unidimensional semantic hierarchy. These tags are user-assigned, and the folksonomies primarily thrive within a specific digital context, namely the social environment. The essence of folksonomies lies in the act of aggregation, which depends on user cooperation. Without this distributed social context facilitating aggregation, tags remain mere isolated words, significant only to the individual user who assigned them (Quintarelli, 2005). See table 1.

Of these four characteristics, Wikipedia's category scheme clearly shares the first one (XXX); it is born in a digital environment. It is worth noting, however, that this is the least idiosyncratic of the four, as, at the present moment, we have many native digital knowledge organization systems (KOS).

In contrast, it does not fulfill the second or the third characteristics at all (See Table 1). It does not fulfill the second one (uncontrolled natural language) because Wikipedia's category scheme does apply vocabulary control techniques, both in the designation of categories and in their structuring. And it does not fulfill the third one (natural language without hierarchy) because those who assign the categories and, naturally, those who establish guidelines, criteria, etc., in its development are editors ("content creators") with varying degrees of authority. This third characteristic necessarily connects to the qualification of the fourth and last characteristic. It is evident that the environment for the creation and assignment of categories on Wikipedia can be described as "social." However, beyond the superficial link, there are profound differences when comparing the decision-making process inherent to folksonomies

	Wikipedia environment	Social Media Environment
1 Digital context	YES	YES
2 Natural Language without control	YES	YES
3 Natural Language without hierarchy	No	YES
4 Tag assignment by users	YES	YES
5 No role differences among collaborators	NO i.e.Role differences between editors and administraors	YES
6 Aggregation of concepts without the ability to correct or delete previous contributions	It is only met in the case of readers or editors. It is not met in the case of administrators. They can initiate deliberation processes to make decisions about category maintenance or deletion.	YES

Table 1. Characterization in 2 different environments of folksonomies. Source: Author's own elaboration.

and tagging and the decision-making process of Wikipedia's category scheme and categorization in this context.

2.2. Description of the categorization scheme

In the previous section, we positioned Wikipedia's KOS in the category schemes and taxonomies in terms of its structure and functions. We ruled out its classification as a folksonomy, despite the distributed mode of construction. In the choice between these two types, we leaned more towards the former (category schemes) than the latter (taxonomies). Now is the time to highlight the distinctions that set Wikipedia's KOS apart from taxonomies, in order to evaluate, through comparison, its idiosyncratic characteristics as a categorization scheme.

For this purpose, we will subject Wikipedia's KOS to a series of key questions about its components and structure. These questions are based on those that ***Guidelines and good practices for taxonomies***, from now on GGPT (I.S.S.T, 2009) considers essential when developing the process of constructing a taxonomy. We group our questions into three dimensions: The identification of concepts to include in the KOS; The classification criteria applied for hierarchy generation; The types of semantic relationships between concepts.

It is in the answers to these questions that we will characterize Wikipedia's categorization scheme, emphasizing two considerations: the features that align and differentiate it from taxonomies and the benefits and risks associated with these different positions, as well as the features that align and separate it from the gender perspective.

2.2.1. Identification of concepts to include in the knowledge oganization system

In this section, we examine the justification for incorporating categories into the Knowledge Organization System (KOS). We also delve into the complexity of these categories, specifically, the number of concepts that are combined to form them. Additionally, we explore the methods used for designating categories.

The guidelines and instructions that support the editors prioritize two criteria in justifying the inclusion of new categories: frequency in articles and usage frequency. At first glance, both criteria are objective. We can observe this in the following references:

- Within Wikipedia:Categorització guidelines (Wikimedia, 2023f) [In English: Categorization guidelines], in the section "When to create a new category," the following rule is established: "In the event that we see that a category has many articles (15 or more) and that some of them (necessarily 5 or more) belong to a more specific area within that branch of knowledge, a new category can be opened to contain these articles. However, we must remember to include the newly created category within the original category from which we took the nominated articles and move the related articles to the new category [...]."
- In the proposed policy Wikipedia:Propuesta de política de categorización (Wikimedia, 2022c) [-in English: Proposal for categorization policy], in the general considerations on categorization, the principle "Do not create categories of little use" is established.

Note that the justification based on the frequency of appearance is established with two specific figures. Categories assigned to 15 or more articles, of which 5 or more are about a more specific concept, constitute the precise threshold for generating a new category.

A second issue is related to the complexity of categories, within Wikipedia's scheme, we find representation of both extremes of the continuum.

- Categories representing a single concept. For example, Infermeria (represents a knowledge area) or Infermers (represents a professional activity, that is to say nursing).
- Categories representing combinations of two or more concepts. For example, Persones de l'àmbit catalanoparlant per origen i activitat (represents the combination of entity + space + origin + activity), Bibliotecaris catalans del sud contemporanis (represents the combination of professional activity (librarian professions)+ origin + time), Infermers catalans del sud contemporanis (represents the combination of professional activity + origin + time), Mestres d'educació primària catalans del sud contemporanis (represents the combination of professional activity (teachers) + origin + time).

The first solution aligns with the post-coordination model, which gives the user the power to select and combine concepts in search and retrieval systems. The second solution, on the other hand, aligns more with the pre-coordination model and has traditionally aimed to facilitate search by exploration and navigation. The application of this second solution has several risks:

- It increases the dimensions of the knowledge organization system, sometimes to unmanageable limits, if there are no constraints on the combinatorial possibilities of concepts. It distances the system from the necessary connection with users' intuition if pre-coordination is not based on users' search frequency.
- It is worth noting, however, that the traditional identification between pre-coordination and systems for exploration and navigation searches cannot be maintained in the context of current digital interfaces, where a sequential journey through concept hierarchies is no longer necessary (or optimal) to reach the term that best satisfies an information need.

Overall, the overuse of pre-coordination diminishes the user's role in decisions regarding concept combinations to meet their needs and expectations in specific moments and over extended periods. On Wikipedia, the absence of guidelines on this issue suggests spontaneous decisions regarding the pre-coordination/post-coordination dilemma and, additionally, a harmful consequence of the "false faceting" phenomenon, which we will discuss later.

A third issue related to the identification of categories to include in the KOS is the designation of categories, which is justified based on the results of Wikipedia's decision-making system. In the case of biographies, the decision made has favored up to 16 categorization criteria (in contrast, as trivial as "Innovators in the motorcycle sector," or as ambiguous as "status," which from a more intersectional perspective, Rodó-Zárate (2021) questions why gender is rejected in category division. This would mean that gender can be used as a general category but not for subcategorization. Currently, in Wikipedia, the general category always uses the masculine gender, except in very specific cases such as Llevadores or Dones Barbudes. Therefore, the default gender in first-level categories is masculine and also in subcategories. This is a consequence of the fact that gender division is not allowed at this level as indicated in the guideline and was corroborated in the vote, except for special cases as mentioned.

The Wikipedia:Categorització guideline from 2018 (before the 2022 gender category vote) justifies this rule based on the lack of consensus and explicitly states the difference in criteria compared to other Wikipedias. The results show that categories related to people apply exclusively the male gender, whether the category directly focuses on a group of people (Científics, Il·lustradors científics, Crítics culturals, Promotors culturals, Directors de tecnologia, all professions writen with male gender) or if the group is part of a more complex designation (Congrés Internacional de Matemàtics). And it is difficult to justify the inclusion of the category Infermers (Male Nurses), which includes seven pages, all related to female individuals (Wikimedia, 2021).

2.2.2. Classification criteria applied for hierarchy generation

Generally, the classification of objects within a KOS can involve various types of hierarchical relationships, each corresponding to distinct logical situations. The standard ***Información y documentación: tesauros e interoperabilidad con otros vocabularios. Parte 1: Tesauros para la recuperación de la información***, from now on UNE-ISO 25964-1 (AENOR, 2014), identifies three such relationships:

- The generic relationship, consisting of the connection between a class or category and its members or species.
- The part-whole relationship, which covers a limited range of situations in which a part of an entity or system belongs exclusively to a particular whole that possesses it. The UNE-ISO 25964-1 standard identifies four of them: systems or organs of the body; geographical locations; disciplines

Feminized profession	Subdivision (original label)	Subdivision (translation)
Nursing (Infermeria)	Nurses Midwives First Aid	Infermers Llevadores Primers Auxilis
Library science (Biblioteconomia)	Bibliography Professional associations related to information and documentation Bibliography Bibliometrics Library of Catalonia Librarians Libraries Documentation centers Directories School of Librarians Faculty of Library and Documentation (UB) IFLA Presidents Information and documentation magazines	Associacions professionals relacionades amb la informació i la documentació Bibliografia Bibliometria Biblioteca de Catalunya Bibliotecaris Biblioteques Centres de documentació Directoris Escola de Bibliotecàries Facultat de Biblioteconomia i Documentació (UB) Presidents de l'IFLA Revistes d'informació i documentació
Primary education (Educació primària)	Students by primary education primary education centers Primary education teachers	Alumnes per centre d'educació primària Centres d'educació infantil i primària Mestres d'educació primària

Table 2. Classification criteria in the subdivision of feminized professions. Source: Author's own elaboration.

or fields of knowledge; hierarchical social structures. Other standards, such as GGPT, substantially expand them.

- The enumerative relationship connects a general concept, such as a class of things or events, with an individual instance of the mentioned class, which is often represented by a proper name.

Regardless of the type of relationship selected, standards and best practices recommend ensuring consistency by generating hierarchy chains. This test consists of the fact that all concepts (categories) from the highest level to the lowest level belong to the same fundamental category, such as objects, materials, people, actions, places, times, etc. It is a guarantee of the semantic disjunction of the represented concepts; that is, there will be no two concepts with semantic overlap.

On the other hand, taxonomies have singled out the generic relationship as the only applicable one, where every object captured by the most general concept is also captured by the more specific concept. This relationship can be expressed as an "is a" relationship; a member of Infermers de Catalunya is a member of Infermers. Applying the generic relationship instead of, for example, the part-whole relationship, requires compliance with an additional logical validity rule; the "all-some" test. All Catalan nurses are nurses; some nurses are Catalan nurses. In this way, the application of the transitive property is guaranteed at all levels of a hierarchical chain. The specific concept is a (member or type) of its general concept; this is a (member or type) of its general concept...; and so on to the broadest concept in the hierarchy chain.

The application of these principles is consistent with the utility expected from the application of hierarchies in information retrieval. On the one hand, the user can navigate from the lowest levels to the highest, or vice versa, with the certainty that they will never lose control over the concepts expected to be found at higher or lower levels. And they will not lose it when the search mode uses KOS for search expansion; that is, the automatic incorporation of specific concepts into the query equation.

The logical tests indicated are not met by other types of relationships, such as the part-whole relationship. And much less those that deal with associations between concepts, such as cause-effect, agent-instrument, area of knowledge-object of study, etc. In all cases, hierarchies generated based on pragmatic criteria, but they depart from intuition and often from the understanding of users.

The classification criteria applied in the Wikipedia categorization scheme do not respond univocally to the generic relationship. In the Spanish version of the Wikipedia categorization guidelines (Wikimedia, 2023e), the lack of this requirement is summarized in the expression "Lack of conceptual transitivity," and is defined as follows: "The main rule is that each subcategory should be more specific than the categories it is included in. However, given the very broad and diverse nature of the categories, conceptual containment may not be inherited by deeper-level subcategories. This does not constitute an error."

The mixing of classification criteria can be clearly observed in the subdivision applied to feminized professions (See Table 2).

The risk of not respecting the generic relationship in establishing the hierarchical relationship between categories is clear: the generated hierarchical chains make it difficult for users to understand the overall structure of the KOS structure, and they are not intuitive and difficult to learn and remember in the process of exploring concepts and expanding searches.

The second question we have raised in this section, about how many categories of broader semantic scope (supercategories) can be linked to a category, leads to the distinction between monohierarchy and polyhierarchy. A KOS is monohierarchical when each concept (except the top concept) has one and only one broader concept. Otherwise, when concepts can have more than one broader concept, we are dealing with a polyhierarchical KOS. In this second case, broader terms are not disjoint, but they overlap. A possible situation in polyhierarchy is when a concept is specific to a superior one based on the generic relationship, and others are generated based on other types of relationships, such as part-whole.

The GGPT classifies monohierarchy as a technique that brings less complexity and, also, less expressiveness to KOS, while polyhierarchy acts in the opposite direction regarding these two values. Polyhierarchy provides various ways to reach the same content, and this diversity is also reflected in the search systems. However, polyhierarchical KOSs tend to be difficult to understand, especially when all possible classifications or distinctions of terms are implemented in parallel.

In Wikipedia's categorization scheme, the application of polyhierarchy is quite common. Furthermore, in cases where it is applied, the number of supercategories is more than two. Examples of this are the categories related to feminized professions (librarians, nursing personnel, and primary and secondary school teaching personnel).

- The Librarians category has the following supercategories: Library Science, Categories with commonscat link from Wikidata, Information Managers, and Professions in Literature.
- The Nurses category has the following supercategories: Biographies by activity, Categories with commonscat link from Wikidata, and Nursing.
- The Primary Education Teachers category has the following supercategories: Categories with commonscat link from Wikidata, Primary Education, and Teachers by Educational Level.

Polyhierarchy is also applied to categories created from the precoordination of more than one concept, as can be seen in the following examples.

- The category People in the Catalan-speaking area by origin and activity (Persones de l'àmbit catalanoparlant per origen i activitat) has the following supercategories: Europeans by origin and activity, People in the Catalan-speaking area by activity, and People in the Catalan-speaking area by origin.
- The category Contemporary South Catalan Librarians (Bibliotecaris catalans del sud contemporanis) has the following supercategories: Catalan Librarians, Spanish Librarians, and Contemporary South Catalans by activity.
- The category Contemporary South Catalan Nurses (Infermers catalans del sud contemporanis) has the following supercategories: Contemporary South Catalans by activity, Catalan Nurses, and Spanish Nurses.

The combination of precoordination and polyhierarchy in the categories mentioned above has a multiplicative effect in terms of their semantic complexity, and in search contexts, it places demands on users for comprehension, learning, and memorization efforts. These demands do not decrease significantly even when polyhierarchy is applied to categories that represent a single concept.

In cases where the same concept can be viewed from different perspectives, as polyhierarchy allows, and there is a need to combine different concepts to represent many of the classified contents comprehensively, an alternative global KOS structuring is faceted classification. Facets are groupings of concepts that represent one and only one feature of the division of an area of knowledge, or of knowledge as a whole. Facets are mutually exclusive, and the set of facets allows for a complete representation of knowledge. There are proposed facets for general application, for all domains of knowledge; Aristotle's 10 fundamental categories, Ranganathan's PMEST formula, or more recently, the 9 types of concepts proposed by the UNE-ISO 25964-1 standard:

1. Objects and their physical parts
2. Materials
3. Activities or processes
4. Events or occurrences
5. Properties of people, things, materials, or actions
6. Disciplines or thematic fields
7. Units of measurement
8. Types of people and organizations
9. Individual entities (represented by proper names)

Wikipedia fulfills three of the recommended scenarios for the application of the faceted classification model according to GGPT as it covers interdisciplinary areas with more than one perspective to view content objects or the need to combine concepts. Second, It has an KOS with multiple hierarchies but unclear boundaries. And third, The KOS is oriented towards the classification of digital objects where location and placement are not important.

And it could benefit from two key advantages highlighted by these guidelines: One advantage, faceted classification has much smaller dimensions than mono- or polyhierarchical ones, without losing its expressiveness. Consequently, it is easier to build and maintain. And the other advantage, faceted classification allows users to search or navigate resources more flexibly, as they can search for a resource from different angles.

Wikipedia's KOS does not make an exhaustive, systematic application of the faceted classification model. This fact keeps it from obtaining the benefits identified by GGPT. There are two main deviations from the strict faceted model.

At the basic level of classification, a proper facet analysis has not been conducted. The eight main thematic categories established definitively in 2016 are not mutually exclusive:

- **Biographies**
- **Science**
- **Culture**
- **Events**
- **Humanities**
- **Information**
- **Locations**
- **Technology**

There are five main categories that respond to the division of human knowledge into disciplines: Science, Humanities, Information, and Technology. However, the criterion applied for the segmentation of disciplines leads to multiple overlaps in the subsequent levels. For example, the category Cultural Studies is an example of this.

In contrast, Events and Locations seem to be generated from fundamental categories, while Biographies directly pertain to a documentary genre. All of these are applicable to all disciplines of human knowledge. In other words, Science can be subdivided into various hierarchical levels of specification by scientific areas or disciplines, but it can also be subdivided by the locations of Science, Science Events, and Science Biographies.

Finally, the Culture category cannot be clearly identified as a discipline or a phenomenon.

In the subdivision of the main thematic categories, we can distinguish three different orientations:

- Subcategorization aimed at subfaceting:
  - **Biographies** (8 categories, 29 pages). They focus on the subcategory "Categories of biographies by parameter."
  - **Events** (9 categories, 5 pages). The facets are detailed in six of the nine subcategories: "Ongoing events," "Events by month," "Events by century," "Events by theme," "Events by territory," and "Events by type."
- Subcategorization aimed at specifying disciplines.
  - **Humanities** (13 categories, 12 pages). Includes: "Administration," "Art," "Law," "Cultural Studies," "Philology," "Philosophy," "Medical Humanities," "Linguistics and Theology."
  - **Information** (5 categories, 1 page). Includes: "Information Sciences," "Communication," and "Journalism."
- The main thematic categories in this third group are characterized by incorporating a large number of subcategories.
  - **Science** (41 categories, 70 pages)
  - **Culture** (32 categories, 54 pages)
  - **Technology** (22 categories, 273 pages)

Among these subcategories, we find, on the one hand, those that could form subfacets, including the discipline subfacet. For example, in the case of Culture, we explicitly find the subfacets Culture by human group and Culture by territory. The other 30 can be simplified as follows.

- Facet of people, where Cultural Critics, Cultural Promoters, etc., would be located.
- Facet of processes and phenomena, where Cultural Anarchism, LGBT Culture, etc., would be located.
- Facet of events, where Cultural Events would be located.
- Facet of objects, where Clothing and Works would be located.
- Facet of disciplines, where Art, Cultural History, and Religion would be located.

The main category Locations exhibits a unique treatment, somewhere between the two previous orientations.

2.2.3. Types of semantic relationships between concepts

The only semantic relationship applied in the Wikipedia category scheme is that of hierarchy, meaning the relationship between a pair of concepts where one has a scope that is completely within the scope of the other. Two other relationships that play a fundamental role in supporting users in exploratory searching are excluded from the structure: the equivalence relationship and the association relationship.

The equivalence relationship is established between two terms when both represent the same concept. The origin of equivalence can be due to natural language synonymy or it can be motivated by identifications between concepts that, even though they are not strictly synonymous in natural language, are treated as such in the context of the system of knowledge organization (KOS) to reduce vocabulary overflow in both the depth and breadth dimensions.

Standards and best practices recommend the incorporation of this semantic relationship in complex KOSs such as Wikipedia's category scheme. However, it is necessary to designate one of the equivalent terms as preferred to prioritize it in contexts where the concurrence of all forms is difficult. From the user's perspective, the equivalence relationship aligns with the visibility of diversity. Different designations of the same concept, often linked to different perspectives, can contribute to the processes of exploration and content retrieval. Additionally, it is possible to implement mechanisms for customizing vocabulary for each user based on their preferences.

Some indications regarding the formal aspects of category designations, although not exhaustive, can be found in Wikipedia: Categorization. Conventions on title designations also extend to categories, with significant specifications. In fact, more explicitly, the Spanish version of Wikipedia: Title Conventions states:"Categories are created analogously to articles, always respecting title conventions; however, all of their content is optional, except for the identification of their parent category or categories."

One of the specificities in category designations is precisely the exclusion of the recommendation to create redirects, which closes the door to the equivalence relationship in Wikipedia's category scheme.

The association relationship is established between a pair of concepts that are not related hierarchically but share a strong semantic connection. Standards and best practices highlight that the incorporation of this relationship enhances the usefulness and expressiveness of a taxonomy by suggesting additional (associated) terms for use in indexing and retrieval, especially named relationships. However, it also entails risks as it increases the complexity of the SKO in its construction and maintenance. In any case, the establishment of associations between concepts cannot be spontaneous or driven by personal views. It is necessary to establish patterns of relationships between concepts to be applied across the entire SKO or in specific sections of the scheme.

4.3. Wikidata's Analysis of the Knowledge Organization System

4.1. The ontologies of Wikidata

Wikidata is a large-scale knowledge base (Vrandečić and Krötzsch, 2014). It is a free, collaborative (a), and multilingual database that includes nearly 300 languages (Kaffee et al., 2017), serving as a secondary database (b) and collecting structured data to support other projects in the Wikiverse, such as Wikipedia (encyclopedia), Wikimedia Commons (repository of images, audios, and audiovisuals), as well as anyone worldwide.

(Piscopo et al., 2017). It is collaborative because human editors input and maintain the data, with contributions coming from both humans and automated bots, either authenticated or anonymously. Human editors establish the rules for content creation and management. Additionally, the platform allows for collaborative editing of its data model by a diverse user community (Piscopo et al., 2017).

And it is secondary because it not only records descriptions of objects and concepts but also the sources that support these descriptions and connections to external databases that complement them.

4.2. Data model

Wikidata is structured through a data model that facilitates the retrieval of its own data. The model chosen by Wikidata is the graph-based data model, more specifically, directed graphs. The smallest unit of a directed graph consists of two nodes connected by an edge. One of the nodes represents an object or concept in the world for which we want to express data. The edge represents an attribute of the first node. And the second node is the value of the attribute described by the edge on the first node.

Each of the two nodes can be linked to other nodes through edges, i.e., properties. It is a very simple structure that allows for expansion without a predefined limit in origin, and in a context of open data, without the constraints of a single corporation.

The graph-based data model has been subject to various standardizations. One of them is the Resource Description Framework (RDF) carried out by the W3C as part of its semantic web project, which aims to provide semantic content to web data and utilize this content to improve the efficiency and effectiveness of navigation and search processes. The RDF data model designates the structure formed by two nodes linked by an RDF statement or triple. The node being described constitutes the subject of the statement, the edge constitutes the predicate (or property of the statement), and the node expressing the value about the subject is the object of the statement.

RDF requires that the subject and the predicate be occupied by resources identified with an Internationalized Resource Identifier (IRI). An IRI is a Unicode string used to uniquely identify nodes and edges. IRIs are internationalized versions of URIs that are generalizations of URLs. Using IRIs helps to avoid naming conflicts and promotes distributed naming, meaning avoiding the need for a centralized naming authority. Additionally, IRIs can be resolvable and accessible via the HTTP protocol, ensuring that the resource and data about the resource are uniquely accessible on the web.

As for the object of the statement, the RDF data model allows it to be occupied either by a resource identified with an IRI (similar to the subject and predicate) or by a literal. The former shares the same characteristics as we have seen regarding the subject and predicate of statements. In contrast, the literal is a concrete value used to represent data types like strings, numbers, dates, etc.

The RDF data model and the entire semantic web project were the origins of Wikidata. However, Wikidata decided not to fully adhere to the requirements of the RDF data model. Nevertheless, interoperability between Wikidata and RDF datasets is well established. Below, we'll discuss the specifics of the data model in Wikidata.

In the Wikidata database, the resource in the subject position is called an "item". For each item, various descriptions are provided in the form of statements or, if viewed from a graph perspective, various labeled edges with nodes pointing to the described items. The set of statements about an item constitutes a page on Wikidata, although they are also available in other machine-readable formats.

Two types of statements are distinguished: basic information about the item to facilitate its identification, designation, and search; and statements with more detailed content about the item. An example of the first type can be seen in Image 11, which corresponds to basic information about the item "Bel Olid Báez (Q16176248)". Basic information includes labels, brief descriptions, and aliases ("also known as") in different languages.

The second type of statements, with more detailed and specific content about each item, constitutes the most important descriptions on Wikidata and, as a result, deserve more attention. Each of these statements consists of two parts: a "claim" that describes some aspect of the item, and a list of references where the claim can be verified. It should be noted that providing references is not a strict requirement for incorporating a claim about an item.

The claim is completed with the property "gender", and the object of this property (its value), which is the non-binary gender. The literal interpretation of this claim is:

"Bel Olid Báez has the property 'gender' with the value of non-binary gender."

The non-binary gender element in the object position of the claim can be the subject of other claims outside the scope of describing Bel Olid Báez (Q16176248).

Regarding this claim, there are two references from sources where its certainty can be verified. The literal interpretation of the first of these references is:

"The claim that Bel Olid Báez has the property 'gender' with the value of non-binary gender is referenced in the source https://twitter.com/BelOlid/status/1491863812311552001. This source was consulted on February 11, 2022."

"The item 'Bel Olid Báez (Q16176248)' has the property 'birth name' with the value of the character string 'Bel Olid Báez.' This character string is expressed in Catalan." In this case, the claim includes a literal as the object, specifically, a character string. Unlike the item in the object position that we saw in the previous claim, the literal "Bel Olid Báez" cannot be the subject of other claims. Its sole purpose is to provide a value for the 'birth name' property related to the item Bel Olid Báez (Q16176248).

Finally, it's worth noting that there are no references verifying it (the value is 0). As is the case with Wikipedia, editors can create claims without references. These claims, just like in Wikipedia, can be improved by other editors who do have references.

The set of claims about an item (basic information, claims, references) forms a web page, the item's web page. In the example regarding the item "Bel Olid Báez (Q16176248)," the web page is located at the URL https://www.wikidata.org/wiki/Q16176248, and it is accessible through a web browser, searchable through web search engines, etc. Beyond the limits of this page, the entity Bel Olid Báez (Q16176248) is identified with a URI, http://www.wikidata.org

4.3. The ontology in the context of Wikidata

Wikidata's ontology is an upper ontology, which means it is applicable to various knowledge domains. From the perspective of its creation and development, it is a "community-controlled ontology." This means that it doesn't start with a predefined group of specialists, nor is it based on an existing one. Currently, the ontology doesn't strictly adhere to the principles of the semantic web, and it is not designed following its standards. However, recently, there has been an effort to align with these standards, and mappings have been created.

The fundamental components of the ontology are properties, which we have already discussed in the context of statements, and classes. A property is a descriptor or attribute for the value of a statement. For example, the "gender" property is the descriptor that assigns the value "non-binary gender" to the item "Bel Olid Báez." A class is a grouping of individual elements defined based on common characteristics among individual elements. For instance, the individual element "Bel Olid Báez" belongs to the "human" class and shares characteristics with all individual elements within that class. In ontology terminology, it is said that "Bel Olid Báez" is an instance of the "human" class.

Furthermore, properties and classes are marked with a series of semantic constraints that establish conditions for their application, and they play a crucial role in more advanced processes of consistency testing and inference. The constraints ensure that the statements generated about an item are semantically correct, meaning they are true. Inference allows for the automatic generation of new statements by applying rules of implication (if A, then B) to explicitly created and published statements in the Wikidata database.

The conditions for creating these components in Wikidata's ontology are detailed below, illustrated within the context of gender, and assessed based on samples and observation. Additionally, we identify areas for improvement that need to be addressed, both for utilizing the ontology within the context of Wikidata and for potential future applications in Wikipedia.

4.3.1. Properties

At Wikidata, users who are recognized with the technical capability can create properties. Administrators also have this right by default. According to the guideline Wikidata:Property creators (Wikimedia, 2023d), there were 116 property creators, of which 65 were administrators.

The conditions for creating a property are detailed in Wikidata:Property creation/es. It is especially significant that the proposal must have been made by someone other than the one who ultimately creates it, and the proposal must have sufficient consensus in favor. This means it must consider the reflection and logic of discussion points. In fact, a well-reasoned dissenting voice can block the creation of the property. Properties are described using the same type of statements we discussed in Section 6.1.1 for Wikidata elements, as exemplified in the case of Bel Olid Báez (Q16176248). However, properties have a very important specificity, which is statements of constraints.

A constraint is a rule on how a specific property should be used. Currently, there are 44 types of constraints, which can be found in Help:Property constraints portal/list of constraints (Wikimedia, 2023c). Each of these types is an element (class) in Wikidata and, like any element, is described through statements. Let's see a couple of examples:

- The "type constraint (Q21503250)" allows specifying what type of entity can be the subject of the property. Entity types correspond to ontology classes.
- The "one-of constraint (Q21510859)" allows specifying that the value of the property must be chosen from a delimited set of elements.

In the Wikidata data model, constraints are declared using triplets composed of the following components:

- The subject of the triplet is the property on which the constraint acts. For example, the "sex or gender (P21)" property.
- The property of the statement is "constraint (P2302)."
- The object of the statement is one of the 44 elements we mentioned before.

The wording of the statements of restrictions related to the "sex or gender (P21)" property is as follows:

- Statement 1: The "sex or gender (P21)" property has a restriction (P2302) of type "type constraint (Q21503250)."
- Statement 2: The "sex or gender (P21)" property has a restriction (P2302) of type "one-of constraint (Q21510859)."

Usually, it is necessary to make statements about these constraint statements to detail aspects such as their scope, whether the constraint is mandatory or optional, and, as we saw in the case of elements, verification references. It is a recursive process where the constraint statement becomes an element itself, so it can occupy the subject position of a statement and be described by other statements.

There are 18 classes whose instances can be the subject of the "sex or gender (P21)" property (See Annex 1). And there are 53 elements whose instances can be the subject of the "sex or gender (P21)" property (See Annex 1).

Wikidata maintains various tools for detecting property constraint violations and facilitating their resolution. However, as we mentioned at the beginning of this section, the correct application is left to the discretion of the editors, so the exploitation of these ontology elements for consistency validation and new knowledge inference processes remains rather limited. This is one of the urgent areas for improvement that ontology stewards have raised, with specific measures that we will describe in the following sections.

4.3.2. Classes and Instances

Wikidata establishes a clear distinction between instances (or individuals) and classes. An instance represents a concept or individual object that is clearly identifiable. And a class is an abstract object representing a set of instances.

As indicated in Help:Basic membership properties (Wikimedia, 2023a), normally, all instances belonging to a class share a set of properties that characterize the class. Instances differ from each other in terms of the values they have for these properties, not in the possession of the properties themselves. Therefore, each class is typically characterized by the properties shared by all its instances (although Wikidata does not strictly enforce this). Wikidata's class item (Q16889133) defines it as a "set of items that share a specific property."

A specificity of Wikidata is the possibility that an item can be both an instance and a class simultaneously. In practice, there is no restriction that prevents an item from being both an instance and a class. According to Wikidata's help page, Wikidata:WikiProject Ontology/Classes (Wikimedia, 2022b), the definition is pragmatic.

- An item is an instance simply because it is the subject of an instance of property (P31). For example, the non-binary gender item (Q48270) is an instance of the gender identity class (Q48264). And the Bel Olid Báez item (Q16176248) is an instance of the non-binary gender class (Q48270).
- An item is a class simply because it is the object of the instance of property (P31) or the subject or object of the subclass of property (P279). For example, the same non-binary gender item (Q48270) is a subclass of gender (Q48277), generic term (Q210588), and gender minority (Q11894636); and a superclass of 30 classes, including kathoey (Q746411) and agender (Q505371). In contrast, the Bel Olid Báez item (Q16176248), representing a unique object in the world, is not a subclass of non-binary gender (Q48270) or any other class.

The sum of statements about items and the statements about the properties and classes that underpin them constitutes Wikidata's knowledge graph.

The instance of (P31) and subclass of (P279) properties are the fundamental edges of the knowledge graph generated by Wikidata components. A third property that complements the graph's structure, albeit subsidiarily, is the part of property. This property can have an instance as its subject, which can be either an instance or a class. For example, the non-binary gender class (Q48270) is part of the trans identity class (Q1771029).

The reference standard in the design and development of classes is based on RDFS (which subsumes the basic RDF model standard), although in a light way.

The instance of (P31) and subclass of (P279) properties correspond, respectively, to the rdf:type and rdfs:subClassOf properties with similar meanings.

- rdf:type = The subject is an instance of a class.
- rdfs:subClassOf = The subject is a subclass of a class.

As in semantic web standards, all instances of a given class are also instances of the superclasses of that class. In other words, since Bel Olid Báez (Q16176248) is an instance of the non-binary gender class (Q48270), it is also an instance of the superclasses gender (Q48277), gender minority (Q11894636), and generic term (Q210588). By ascending the hierarchical chains of these three classes, we arrive at the final statement that Bel Olid Báez (Q16176248) is an instance of the root class of Wikidata's ontology, ens (Q35120), described as "something that exists."

However, the adherence to semantic web standards is "light," as indicated in Wikidata:Item classification. One manifestation of this lightness is the promiscuity between instances and classes that we referred to.

While Wikidata's ontology does not adhere to the OWL (Web Ontology Language) standard, some semantic restrictions are applied to classes' creation, similar to those in OWL. For example, the subclass of (P279) property is an instance of the transitive property class (Q18647515). According to this definition, if A is a subclass of B, and B is a subclass of C, it follows that A is a subclass of C. This axiom allows the generation of class hierarchies from the top node (more general entities) to the lower nodes (more specific entities), which is useful for exploration or query-based navigation in certain contexts.

This inference can be generated in environments where statements are accompanied by effective inference rules.

Wikidata incorporates some instruments in this line, including:

- Class Entity Fundamental of Wikidata (Q115490628), which currently includes 11 direct instances, such as the transitive property class (wd:Q18647515).
- Class Property of Wikidata for the relationship between classes (Q28326461), which currently includes 5 direct instances, including the subclass of property (P279).

However, the number of class restriction definitions for generating inferences and controlling consistency with their resources is limited, incomparably lower than what semantic web standards offer.

Unlike properties, the creation of classes can be undertaken by any editing individual. This openness likely enriches the perspectives with which Wikidata's ontology is created, but it also carries its risks, as we will see in the evaluation based on the indicators presented below.

The developers of Wikidata's ontology and the entire knowledge graph are aware of the need for improvements in the ontology creation process. Different resources have been created to facilitate the use of the ontology in editing and entity enrichment. Some of these resources are oriented towards consultation while editing or enriching entities. Examples include Help:Property constraints portal (Wikimedia, 2023b), which presents rules on how properties should be used, and lists of properties recommended by experts for describing specific entity types; for example, Wikidata:WikiProject Biography (Wikimedia, 2022a) for describing biographies.

However, there is a resource that guides the task of editing specific entity categories and prescribes the types, properties, and values to be assigned. These are the Entity Schemas, formally expressed using a specific language (Shape Expression language or ShEx). They have been created for an increasing number of entity types, including individuals. For the creation of this category of entities, we currently have two schemas:

- Entity Schema E10, used to describe instances of the human class (Q5) through 22 properties.
- Entity Schema E14 (basic properties), used to describe instances of the human class in a simpler way, using only six properties.

It's important to note that both schemas incorporate the property gender or sex (P21) prescriptively.

For the application of these schemas, editors have access to tools for generating items from data in CSV format, entity validators generated from schemas, and more. Overall, this project aims to improve the quality of published data and the usability of this data in search contexts, while also fostering the growth of the editor community through simplifying processes and reducing conflicts.

4.3.3. Labelling of items and properties

Wikidata's items and properties have expressions in natural language, collectively referred to as terms. There are three types of terms: labels, which are the primary designations for identifying entities; aliases, which are secondary designations; and descriptions, which are explanatory texts about the entities. Wikidata's goal is to make these components available in all languages.

Labels are especially relevant in search contexts. There can only be one label per language. They serve as the communication tool between the user's needs and the entities that make up the ontology. The success of user search processes in various modalities (navigation, exploration, and querying) depends on the precision and comprehensiveness with which entities have been designated.

User Mr. Ibrahem (a Wikidata user) publishes statistics on the percentage of item labels available by language. As of March 6, 2023 (the last record at the time of writing this report), the percentage of labels in the Catalan language was 14.9% (in absolute figures, 15,116,853 out of 101,045,411 possible). At that time, it ranked eighth out of 539 languages. It's worth noting that while the first two languages, English and Dutch, achieve notable availability percentages, 85.8% and 62.3%, respectively, starting from the third position, the availability percentage drops to 24.2%.

On the other hand, the situation for properties is much more favorable for the Catalan language. As of June 29, 2023, there were 11,051 properties on Wikidata, and of these, 98.20% were available in Catalan (in absolute figures, 10,853).

Another relevant aspect of labels in the context of this report is their alignment with a gender perspective. Property creators and editors have a help page, Help:Label/ca (Wikimedia, 2016), with a warning that it is a translation of proposals for English and only the "starting point for future label guidelines in Catalan." The original English version does incorporate a guideline aligned with a gender perspective. "If the label is different in the male form and in the female form, it is recommended to use a gender-neutral form if it exists in common usage (i.e., use a form that applies to males and females). Avoid using the male form as a generic form when possible. Examples: bomber (Q107711): firefighter is gender-neutral, and fireman is a male form. It is better to use firefighter."

In contrast, in the Catalan version, this guideline has not been included in the translation. In fact, labels formed by names derived with the suffix -tor/-tora only adopt the male form: autor, compositor, director, editor, elector, gerent/director, locutor, productor, rector, traductor... and in all compound terms with these words.

These features of labels and, specifically, those generated for property designations have significant consequences for Wikidata's search services, as we will see in the following sections.

5. Conclusions

The evaluation of Wikipedia categories based on knowledge organization system standards has revealed significant opportunities for improvement, particularly in the realm of gender identity and the broader knowledge organization system. The inconsistencies in the treatment of gender-related categories and the acceptance of top-level categories in the feminine form underscore the need for a more comprehensive and inclusive approach to knowledge categorization. This evaluation suggests that there may not be an objective criterion for rejecting gender identities as categorization criteria on Wikipedia. and the incorporation of gender as a classification criterion is not only justifiable but also aligned with Wikipedia's principles and foundations.

Furthermore, the recognition of gender identity as a criterion for knowledge organization is not without precedent. Libraries, which have long been guardians of universal knowledge access, have successfully integrated gender as a criterion into their organizational systems. This approach has been widely adopted across various language editions of Wikipedia, with the exception of a few.

In the context of Wikidata, its ontology represents a powerful tool with significant potential for user-oriented search systems and a commitment to a gender perspective. The factual nature of the data and its framework of representation allow for the objectification of gender, drawing from external and corroborated sources.

The custodians of Wikidata's ontology are actively addressing its weaknesses and are in the process of implementing corrective measures. This includes the development of data editing and enrichment tools that aim to systematize and maintain consistency in properties, classes, and labels. The Entity Schemas tool is a noteworthy example of such efforts.

Looking ahead, there is an anticipation of embracing Semantic Web languages and technologies, along with the adoption of standards for controlling ontology consistency and inferring new knowledge. Among these standards, SHACL (Shapes Constraint Language), which allows validating RDF graphs against a set of conditions, stands out as a valuable resource for creating and validating RDF graphs.

For further research, it is recommended to apply the same methodology to other editions of Wikipedia, taking into account variations in the treatment of gender in grammar. Additionally, ongoing work includes a user study incorporating usability tests focusing on navigation and search functionalities.

References

AENOR. (2014). *UNE-ISO 25964-1: Información y documentación: Tesauros e interoperabilidad con otros vocabularios. Parte 1: Tesauros para la recuperación de la información*. AENOR. https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0053960

Dawe, L. & Robinson, A. (2017). Wikipedia editing and information literacy: A case study. *Information and Learning Science, 118*(1/2), 5–16. https://doi.org/10.1108/ILS-09-2016-0067

Ferran-Ferrer, N., Castellanos-Pineda, P., Minguillón, J. & Meneses, J. (2021). The gender gap on the spanish wikipedia: Listening to the voices of women editors. *Profesional de La Informacion, 30*(5), e300516. Scopus. https://doi.org/10.3145/epi.2021.sep.16

I.S.S.T, Fraunhofer. (2009). *Guidelines and good practices for taxonomies* (1.3). Semantic Interoperability Centre Europe. https://joinup.ec.europa.eu/sites/default/files/document/2011-12/guidelines-and-good-practices-for-taxonomies-v1.3a.pdf

Kaffee, L-A., Piscopo, A., Vougiouklis, P., Simperl, E., Carr, L. & Pintscher, L. (2017). A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. *Proceedings of the 13th International Symposium on Open Collaboration*, 1–5. https://doi.org/10.1145/3125433.3125465

Maciá, Y. (2022). *Mujeres de categoría: Utilización de los principios y estándares de los datos enlazados (linked open data) para visualizar las biografías de mujeres en la Viquipèdia* [Master's Degree, Universitat de Barcelona]. https://diposit.ub.edu/dspace/handle/2445/189877

Piscopo, A., Phethean, C. & Simperl, E. (2017). What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.), *Social Informatics* (pp. 305–322). Springer International Publishing. https://doi.org/10.1007/978-3-319-67217-5_19

Quintarelli, E. (2005). Power to the people. *ISKO Italy-UniMIB Meeting, Milan, June 24, 2005*.

Singer, P, Lemmerich, F, West, R, Zia, L, Wulczyn, E, Strohmaier, M. &Leskovec, J. (2017). Why We Read Wikipedia. *Proceedings of the 26th International Conference on World Wide Web*, 1591–1600. https://doi.org/10.1145/3038912.3052716

Soler-Adillon, J, Pavlovic, D. & Freixa, P. (2018). Wikipedia in higher education: Changes in perceived value through content contribution. *Comunicar, 26*(54), 39–48. https://doi.org/10.3916/C54-2018-04

Vrandečić, D. & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM, 57*(10), 78–85. https://doi.org/10.1145/2629489

Wikimedia. (2016). *Help:Label/ca*. https://www.wikidata.org/wiki/Help:Label/ca

Wikimedia. (2021). Categoria:Infermers. In *Viquipèdia, l'enciclopèdia lliure*. https://ca.wikipedia.org/w/index.php?title=Categoria:Infermers&oldid=26858616

Wikimedia. (2022a). *Wikidata:WikiProject Biography*. https://www.wikidata.org/wiki/Wikidata:WikiProject_Biography

Wikimedia. (2022b). *Wikidata:WikiProject Ontology/Classes*. https://www.wikidata.org/wiki/Wikidata:WikiProject_Ontology/Classes

Wikimedia. (2022c). Wikipedia:Propuesta de política de categorización. In *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Wikipedia:Propuesta_de_pol%C3%ADtica_de_categorizaci%C3%B3n&oldid=142571261

Wikimedia. (2023a). *Help:Basic membership properties*. https://www.wikidata.org/wiki/Help:Basic_membership_properties

Wikimedia. (2023b). *Help:Property constraints portal*. https://www.wikidata.org/wiki/Help:Property_constraints_portal

Wikimedia. (2023c). *Help:Property constraints portal/list of cons-*

traints. https://www.wikidata.org/wiki/Help:Property_constraints_portal/list_of_constraints

Wikimedia. (2023d). *Wikidata:Creators de proprietats*. https://www.wikidata.org/wiki/Wikidata:Property_creators/oc

Wikimedia. (2023e). Wikipedia:Categorización. In *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/w/index.php?title=Wikipedia:-Categorizaci%C3%B3n&oldid=152512537

Wikimedia. (2023f). Viquipèdia:Categorització. In *Viquipèdia, l'enciclopèdia lliure*. https://ca.wikipedia.org/w/index.php?title=Viquip%-C3%A8dia:Categoritzaci%C3%B3&oldid=32543623

Yedid, N. (2013). Introducción a las Folksonomías: Definición, Características y Diferencias con los Modelos Tradicionales de Indización. *Información, cultura y sociedad, 29*, Article 29. http://revistascientificas.filo.uba.ar/index.php/ICS/article/view/673

Yin, R. K. (2009). *Case study research: Design and methods*. Sage.

Zeng, M. L. (2008). Knowledge Organization Systems (KOS). *Knowledge Organization, 35*(2–3), 160–182. https://doi.org/10.5771/0943-7444-2008-2-3-160

Zhang, C. C. & Terveen, L. (2021). Quantifying the Gap: A Case Study of Wikidata Gender Disparities. *17th International Symposium on Open Collaboration*, 1–12. https://doi.org/10.1145/3479986.3479992

CV

Miquel Centelles

- miquel.centelles@ub.edu
- https://orcid.org/0000-0003-1739-4889
- He is a professor at the Faculty of Information and Audiovisual Media at the University of Barcelona (FIMA). He holds a degree in Library Science and Documentation and a bachelor's degree in Philology. His teaching and research focus on the representation and organization of information, as well as the application of semantic technologies in information and knowledge management. He coordinated the Master's in Digital Content Management from 2005 to 2008, and since 2020, he has been the coordinator of the Master's in Digital Humanities, involving five faculties at the UB. In research, he has collaborated on the Archiver project for the digital preservation of research data (Archiver TENDER – European Union), and the I+D+I project, Women and Wikipedia (PID2020-116936RA-I00).

Núria Ferran-Ferer

- nferranf@ub.edu
- https://orcid.org/0000-0002-9037-8837
- An associate professor at the Faculty of Information and Audiovisual Media at the University of Barcelona (UB) since 2021, previously at the UOC from 2005. She coordinates the Doctoral Program in Information and Communication at UB. She holds a European doctorate from UB (2010), with degrees in Journalism (UAB, 1998), Documentation (UOC, 2003), and a Master's in Information Society (IN3-UOC, 2005). She has been an associate professor at several universities, including UAB, UB, and UPF. Currently, she serves as the delegate of the rector for the Directorate of the Equality Unit. In research, she is the principal investigator of the I+D+I project, Women and Wikipedia (PID2020-116936RA-I00), where she supervises two theses. She has also collaborated on open science and citizen science projects and conducted research stays at the University of Sheffield (United Kingdom, 2009) and the University of Tallin (Estonia, 2015).

Annex 1

There are **18 classes** whose instances can be the subject of the "sex or gender (P21)" property:

1 person
2 animal
3 character (whether fictional or not)
4 abstract entity
5 fictional character
6 mythological entity
7 alter ego
8 fossil
9 organism
10 robot
11 sex doll
12 synthetic voice
13 taxon
14 kunya
15 fetus
16 stillborn infant
17 doll or action figure model
18 fictional creature

There are **53 elements** whose instances can be the subject of the "sex or gender (P21)" property.

1 male
2 female
3 intersex
4 transgender woman
5 transgender man
6 non-binary
7 fa'afafine
8 māhū
9 kathoey
10 fakaleitī
11 hijra
12 masculine
13 feminine
14 unknown value
15 no value

16 two-spirit
17 transmasculine
18 transfeminine
19 muxe
20 intersex organism
21 agender
22 genderqueer
23 gender fluid
24 neutral
25 eunuch
26 pangender
27 co-genitor
28 neutral sex
29 hermaphroditism
30 cisgender woman
31 cisgender man
32 third sex
33 gender X
34 demiboy
35 demigirl
36 bigender
37 transgender person
38 transvestite
39 'akava'ine
40 assigned female at birth
41 assigned male at birth
42 androgynous
43 yinyang ren
44 boi
45 intersex person
46 gynandromorph
47 Takatāpui
48 undisclosed gender
49 Fakafifine
50 fakafafine
51 gender determined by the player
52 gender not disclosed in work
53 androgynos